

Evaluation of a training program intended to calibrate examiners of temporomandibular disorders

Lars Dahlström, Stephen D. Keeling, James R. Fricton,
Susan Galloway Hilsenbeck, Gary M. Clark and John D. Rugh

Public Dental Service, Mölndal Hospital, Mölndal, Sweden, and Department of
Orthodontics, University of Florida, Gainesville, Florida; Department of Diagnostic and
Surgical Sciences, University of Minnesota, Minneapolis, Minnesota; and Department of
Medicine and Department of Orthodontics, University of Texas Health Science Center,
San Antonio, Texas; USA

Dahlström L, Keeling SD, Fricton JR, Galloway Hilsenbeck S, Clark GM, Rugh JD.
Evaluation of a training program intended to calibrate examiners of temporomandibular
disorders. *Acta Odontol Scand* 1994;52:250–254. Oslo. ISSN 0001–6357.

The adequacy of a training program intended to calibrate examiners of temporomandibular disorders (TMD) was evaluated. Eight examiners blindly rated 12 subjects with various TMD signs and symptoms after participating in a 5-h intense training procedure. Some examiners had earlier experience of TMD examinations, and some were newly trained. The Cranio-mandibular Index was used as the assessment instrument. Agreement within the whole group of examiners was low. Training tended to increase the probability of correctly registering signs. It is concluded that the program was not sufficient to create reliability among multiple examiners. More extensive training, not only to a standard but also between the different examiners, appears necessary. Revision of the examination technique is suggested, and recommendations for strengthening the calibration procedures are made. □ *Diagnosis; stomatognathic system; temporomandibular joint syndrome*

Lars Dahlström, Public Dental Service, Mölndal Hospital, S-431 80 Mölndal, Sweden

Evaluation of temporomandibular disorders (TMD) relies principally on reports of subjective symptoms, history, and clinical examinations. There is agreement that the cardinal signs of TMD include limitation of jaw movements, joint sounds, and tenderness to palpation of the masticatory and associated muscles and temporomandibular joints (TMJ) (1, 2). These characteristics are commonly used as diagnostic criteria and for primary classification in clinical practice and research.

The importance of accurate and reproducible assessment methods is obvious, but it is difficult to achieve this with precision in TMD (3, 4). To be interpreted with confidence, clinical observations must be reliably measured by different examiners and over time.

Most TMD clinical signs show low constancy over time (5). This may be due to actual changes in the condition of the patient or to poor reproducibility of the evaluation

methods used. By observing the constancy of clinical TMD signs without the influence of time, it has been concluded that the between-observer variation may be considerably greater than the within-observer variation (6–8).

Reproducibility also seems to depend on which sign is assessed and several other factors. Palpation to assess tenderness in muscles or joints has low reproducibility, whereas assessing deviation in range of motion has a moderate to high reproducibility. Reproducibility is usually highest in assessment of mandibular movements (6–13). Reproducibility may also partly depend on the severity of the clinical signs; for example, if only reflex responses to palpation are noted, reproducibility increases (5, 10). For this reason, it is important to include subjects with different degrees of clinical signs when evaluating observer agreement (11). Experience with TMD examinations per se may not necessarily increase reliabil-

ity, but co-training (14), standardization (7, 15), and calibration (11) should.

In summary, previous studies have shown that reliability between examiners depends on several factors: the patients observed, the methods used, the variables studied, and the training and the characteristics of the examiner. Critical issues for reliability studies in the context of TMD have recently been reviewed (16). Careful definitions of variables, standardization, co-training, and calibration are of importance to increase reliability, but to an unknown extent. The aim of this investigation was to develop and test the adequacy of an examiner training program for clinical trials. This included a test to determine the between-observer reliability in assessing TMD signs. In a post hoc analysis the effects of the examiner's experience and previous training were also evaluated.

Materials and methods

Eight examiners—six dentists, one dental assistant, and one dental hygienist—participated as examiners. Seven of the examiners were instructed and trained in all examination variables and scoring systems by the eighth, a very experienced examiner (J. R. Friction) who served as the trainer and standard examiner. The training–calibration session included watching a 19-min instruction videotape prepared by the standard examiner, clinical demonstration of the examination technique by the standard examiner (40 min), and examination of the standard examiner by the trainees and corrections (60 min). The session also included a discussion of operational definitions of the recorded variables and evaluation procedures and use of the scoring system (60 min). Five test patients were also examined by the seven examiners, supervised by the standard examiner. The trainees' findings were compared with the trainer's findings (120 min). Differences in techniques and scorings were discussed and corrections made. The total training procedure lasted approximately 5 h and was followed, after a lunch break, by the blinded calibration trial with 12 subjects.

The Craniomandibular Index (CMI) (7) was used as the assessment instrument. The CMI consists of 62 categorical items. It provides an overall dysfunction score and two subscales. The evaluation of mandibular movements, TMJ noise, and tenderness of the TMJ are summarized in the Dysfunction Index (DI), and the tenderness of masticatory and associated muscles is summarized in the Muscle Index (MI). The CMI and each subscale may each vary from 0 (no signs positive) to 1.0 (all signs positive).

Four of the seven examiners had previously been co-trained in this particular examination method and are identified as the previously trained examiners (PTE). They had been carefully instructed and trained and had been calibrated with the standard examiner approximately 1 year previously. The PTE had been using the CMI regularly for several months. Of the other three examiners, some had experience with TMD examinations, but none had been trained formally in the CMI examination technique and are identified as the newly trained examiners (NTE).

Six of the 12 examined subjects were patients at the University of Texas Health Science Center at San Antonio (UTHSCSA) Facial Pain Clinic, referred for various signs and symptoms of TMD. The remaining six subjects had not sought care for facial pain and did not report complaints of facial pain. Three were oral implant patients of the Department of Oral Surgery, and three were staff at UTHSCSA, unknown to the examiners. Ten of the 12 subjects were women; the mean age for the group was 36 years (range, 20–62 years). All were compensated (US\$35) for their participation.

The subjects were examined in random order by the eight examiners independently, and each examiner and subject were blind to the other examiners' scores. The examination lasted for about 10 min for each subject and was done in dental chairs in a bay of enclosed cubicles at UTHSCSA.

To measure consistency between examiners, product-moment correlation coefficients and intraclass correlation coefficients (ICCs) were used for continuous CMI, DI, and MI scales (17), and kappa statistics

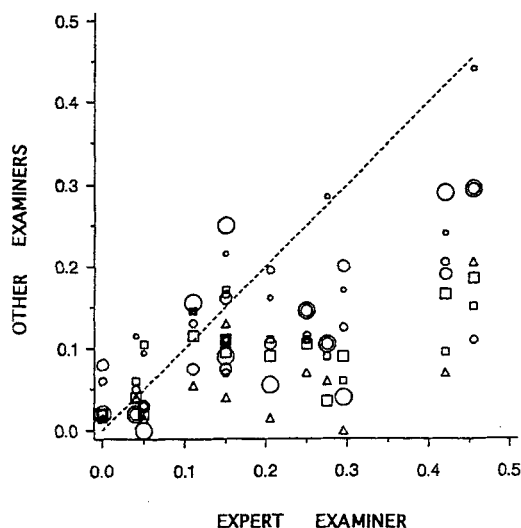


Fig. 1. Scatter plot of Craniomandibular Index scores for the standard examiner and the other 7 examiners for all 12 subjects. The previously trained examiners are indicated by circles of various sizes. Angled symbols indicate newly trained examiners.

(Cohen's kappa) were calculated for each of the categorical variables (18). The kappa statistic adjusts for the likelihood of agreement by chance. Analysis of variance with repeated measures in patients was used to compare examiners' average scores.

Results

The CMI scores for the 12 individual subjects, as evaluated by the 8 individual examiners, ranged from 0 to 0.45. Several subjects without subjective complaints were found to have some signs of TMD.

A scatter plot of CMI scores of the standard examiner and the seven individual examiners is presented in Fig. 1. The ICC for the eight examiners for the total CMI was 0.44. The ICCs for the DI and MI subscales were 0.35 and 0.43. When only the four PTE were considered, the ICCs were 0.54 for the total CMI, 0.43 for the DI, and 0.51 for the MI. When only the three NTE were considered, the ICCs were 0.51 for the total CMI, 0.28 for the DI, and 0.58 for the MI. One of

Table 1. Correlations (product-moment correlation coefficients) between the standard examiner and the four individual previously trained examiners (PTE) and the three individual newly trained examiners (NTE) for the Craniomandibular Index (CMI) and the subscales Dysfunction Index (DI) and Muscular Index (MI)

Examiner	CMI	DI	MI
PTE 1	0.82*	0.61	0.82*
PTE 2	0.57	0.65	0.37
PTE 3	0.87*	0.80*	0.77*
PTE 4	0.71*	0.35	0.79*
NTE 1	0.55	0.09	0.56
NTE 2	0.32	-0.03	0.50
NTE 3	0.78*	0.53	0.71*

Correlations (*) are significant ($P < 0.01$).

the previously trained examiners and the standard examiner had an ICC for CMI of 0.75. All other ICCs between individual examiners and the standard examiner for the CMI or its subscales were lower.

The product-moment correlation coefficients between individual examiners and the standard are presented in Table 1. Three of the four PTE had statistically significant correlations with the standard in CMI, three in MI, and one in DI. Only one of the three NTE had a statistically significant correlation with the standard examiner on CMI and in MI. None correlated well on DI.

The reliability of the components of the CMI assessment was also determined. Grouping of items relating to mandibular movement in the CMI resulted in an ICC of 0.29 for all examiners. The items for TMJ noise showed an ICC of 0.46. Grouping items for extraoral muscle palpation resulted in an ICC of 0.23, the items for neck muscles resulted in an ICC of 0.40, and the items for intraoral muscle palpation resulted in an ICC of 0.38. The corresponding figures for the PTE ranged from 0.21 to 0.45 and from -0.31 to 0.65 for the NTE.

Kappa statistics were used to evaluate the agreement between examiners for the 62 individual items in the CMI. Following the suggestions of Morris et al. (19) for interpretation of kappa statistics, 11 of the items had a fair agreement (0.21-0.40), one had a moderate (0.41-0.60), and none had sub-

